

Approximate Joint Matrix Triangularization

Nicolo Colombo*

Nikos Vlassis†

July 5, 2016

Abstract

We consider the problem of approximate joint triangularization of a set of noisy jointly diagonalizable real matrices. Approximate joint triangularizers are commonly used in the estimation of the joint eigenstructure of a set of matrices, with applications in signal processing, linear algebra, and tensor decomposition. By assuming the input matrices to be perturbations of noise-free, simultaneously diagonalizable ground-truth matrices, the approximate joint triangularizers are expected to be perturbations of the exact joint triangularizers of the ground-truth matrices. We provide *a priori* and *a posteriori* perturbation bounds on the ‘distance’ between an approximate joint triangularizer and its exact counterpart. The *a priori* bounds are theoretical inequalities that involve functions of the ground-truth matrices and noise matrices, whereas the *a posteriori* bounds are given in terms of observable quantities that can be computed from the input matrices. From a practical perspective, the problem of finding the best approximate joint triangularizer of a set of noisy matrices amounts to solving a nonconvex optimization problem. We show that, under a condition on the noise level of the input matrices, it is possible to find a good initial triangularizer such that the solution obtained by any local descent-type algorithm has certain global guarantees. Finally, we discuss the application of approximate joint matrix triangularization to canonical tensor decomposition and we derive novel estimation error bounds.

1 Introduction

We address an estimation problem that appears frequently in engineering and statistics, whereby we observe noise-perturbed versions of a set of jointly decomposable matrices M_n , and the goal is to recover (within a bounded approximation) some aspects of the underlying decomposition. An instance of this problem is *approximate joint diagonalization*:

$$\hat{M}_n = M_n + \sigma W_n, \quad M_n = V \text{diag}([\Lambda_{n1}, \dots, \Lambda_{nd}])V^{-1}, \quad n = 1, \dots, N, \quad (1)$$

where \hat{M}_n are the $d \times d$ observed matrices, and the rest of the model primitives are unobserved: $\sigma > 0$ is a scalar, W_n are arbitrary noise matrices with Frobenius norm $\|W_n\| \leq 1$, and the matrices V, Λ define the joint eigenstructure of the ground-truth matrices M_n . The optimization problem involves estimating from the observed matrices \hat{M}_n the eigenvalues Λ and/or the common factors V . Joint matrix diagonalization appears in many notable applications, such as independent component analysis (Cardoso and Souloumiac, 1996), latent variable model estimation (Balle et al., 2011; Anandkumar et al., 2014), and tensor decomposition (De Lathauwer, 2006; Kuleshov et al., 2015).

Under mild conditions, the ground-truth matrices M_n in (1) can be *jointly triangularized*, which is known as the (real) joint or simultaneous Schur decomposition (Horn and Johnson, 2012). Namely, there exists an orthogonal matrix U_\circ that simultaneously renders all matrices $U_\circ^T M_n U_\circ$ upper triangular:

$$\text{low}(U_\circ^T M_n U_\circ) = 0 \quad \text{for all } n = 1, \dots, N, \quad (2)$$

where $\text{low}(A)$ is the strictly lower-diagonal part of A defined by $[\text{low}(A)]_{ij} = A_{ij}$ if $i > j$ and 0 otherwise. On the other hand, when $\sigma > 0$ the noisy matrices \hat{M}_n in (1) cannot be jointly triangularized exactly. The problem of *approximate joint triangularization* can be defined as the following optimization problem over the manifold of orthogonal matrices $\mathbb{O}(d)$:

$$\min_{U \in \mathbb{O}(d)} \mathcal{L}(U) = \sum_{n=1}^N \|\text{low}(U^T \hat{M}_n U)\|^2. \quad (3)$$

In words, we are seeking an orthogonal matrix U such that the matrices $\hat{T}_n = U^T \hat{M}_n U$ are approximately upper triangular. This is a nonconvex problem that is expected to be hard to solve to global optimality in general. When $\sigma > 0$, the global

*LCSB, University of Luxembourg

†Adobe Research, San Jose, CA

minimum of $\mathcal{L}(U)$ will not be zero in general, and for any feasible $U \in \mathbb{O}(d)$ some of the entries below the main diagonal of each \hat{T}_n may be nonzero. The estimands of interest here could be the joint triangularizer U and/or the approximate joint eigenvalues on the diagonals of \hat{T}_n .

Applications of (approximate) joint matrix triangularization range from algebraic geometry (Corless et al., 1997), to signal processing (Haardt and Nossék, 1998), to tensor decomposition (Sardouie et al., 2013; Colombo and Vlassis, 2016). When the ground-truth matrices M_n are symmetric, the models (1) and (2) are equivalent and V, U_o are both orthogonal. However, when the matrices M_n are non-symmetric, the matrix V in (1) is a general nonsingular matrix, while the matrix U_o in (2) is still orthogonal. Since the optimization in (3) is over a ‘nice’ manifold, approximate joint triangularization is expected to be an easier problem than approximate joint diagonalization, the latter involving optimization over the manifold of invertible matrices (Afsari, 2008). Two types of methods have been proposed for optimizing (3), Jacobi-like methods (Haardt and Nossék, 1998), and Newton-like methods that optimize directly on the matrix manifold $\mathbb{O}(d)$ (Afsari and Krishnaprasad, 2004; Colombo and Vlassis, 2016). Both methods are of iterative nature and their success depends on a good initialization.

1.1 Contributions

We are interested in theoretical guarantees for solutions U computed by *arbitrary* algorithms that optimize (3). Note that the objective function (3) is continuous in the parameter σ . This implies that, for σ small enough, the approximate joint triangularizers of \hat{M}_n can be expected to be perturbations of the exact triangularizers of M_n . To formalize this, we express each feasible matrix U in (3) as a perturbation of an exact triangularizer U_o of the ground-truth matrices M_n in (1), that is

$$U = U_o e^{\alpha X}, \quad \text{where } X = -X^\top, \quad \|X\| = 1, \quad \alpha > 0, \quad (4)$$

where X is a skew-symmetric matrix and e denotes matrix exponential. Such an expansion holds for any pair U, U_o of orthogonal matrices (see for example Absil et al. (2009)). The scalar α in (4) can be interpreted as the ‘distance’ between U and U_o .

Perturbation bounds. We provide two different types of bounds on the parameter α : *A priori* bounds that are based on ground-truth quantities (such as the ground-truth matrices, the sample size, and in some applications also the assumed probability distribution generating the data), and *a posteriori* bounds that involve solely observable quantities (such as the observed matrices and the current solution). While the former bounds are attractive theoretically as they can capture general perturbation effects on the matrix decomposition factors, the latter bounds can have more practical use, such as for instance in nonconvex optimization (Pang, 1987) and the design of optimized algorithms (Prudhomme et al., 2003).

A priori analysis: In Theorem 1 and Theorem 2 we provide two bounds that together offer a complete first-order characterization of the approximate triangularizers in terms of ground-truth quantities. The corresponding inequalities depend on the noise level, the condition number of the joint eigenvectors matrix, a joint eigengap parameter, the number of ground-truth matrices, and their norm. Theorem 2 is the extension of the result derived by Cardoso (1994) for symmetric matrices.

A posteriori analysis: In Theorem 3 we provide an error bound on the perturbation parameter α , which is based on observable quantities that can be computed from the input matrices \hat{M}_n . In particular, the bound involves the value of $\mathcal{L}(U)$ evaluated at each candidate U , and various functions of the observed matrices \hat{M}_n and their approximate joint eigenvalues. The only non-observable quantity appearing in the bound is the noise parameter σ in (1), which, for instance in the case of joint matrix decomposition problems arising from empirical moment matching (see, e.g., Anandkumar et al. (2014)), can be bounded by a function of the sample size. The bound in Theorem 3 is global, in the sense that it does not depend on the initialization, and can be used to characterize the output of any algorithm that optimizes (3).

Global guarantees for locally convergent algorithms. Beyond the purely theoretical analysis of approximate joint matrix triangularization, we also address the practical problem of computing an approximate joint triangularizer in (3). Due to the nonconvexity of (3), locally convergent algorithms are guaranteed to converge to a given local optimum if the algorithm is started in the corresponding basin of attraction. The continuity in the parameter σ of the objective function $\mathcal{L}(U)$ in (3) can be used to show that, under certain conditions, a finite set of local minimizers of (3) enjoy global success guarantees in terms of their distance to the ground-truth matrices. In Theorem (4) we provide a condition under which it is always possible to initialize a locally convergent algorithm in the basin of attraction of such a provably good minimizer.

1.2 Related work

The problem addressed here has two main antecedents: The work of Konstantinov et al. (1994) on the perturbation of the Schur decomposition of a single matrix, and the work of Cardoso (1994) on the perturbation of joint diagonalizers. Our analysis can be viewed as an extension of the analysis of Konstantinov et al. (1994) to the multiple matrices case, and an extension of the analysis of Cardoso (1994) to joint matrix triangularization. We note that joint matrix triangularization is equivalent to joint spectral decomposition when the commuting matrices are symmetric. The proof of Theorem 2 exploits the same idea of Cardoso (1994), but with a few key technical differences that pertain to non-symmetric / non-orthogonal matrices. We are not aware of other works dealing with the perturbation of joint matrix triangularizers. Moreover, to the best of our knowledge, our bound in Theorem 3 is the first *a posteriori* error bound for joint matrix decomposition problems.

From an algorithmic point of view, various approaches to approximate joint matrix triangularization have been proposed in the literature. The simplest one is a matrix-pencil technique (see for example Corless et al. (1997)) where a linear combination of the input matrices is decomposed using established methods for the Schur decomposition of a single matrix. The solution obtained in that case is, however, not optimal and depends on the particular matrix pencil. A more standard way to formulate an approximate joint decomposition problem is to introduce a nonconvex objective function, as in (3), whose variables are the target shared matrix components (Cardoso and Souloumiac, 1996; Haardt and Nossek, 1998; Abed-Meraim and Hua, 1998; Fu et al., 2006; Kuleshov et al., 2015). The nonconvex optimization problem is then solved via iterative methods that typically belong to two classes, Jacobi-like methods (Cardoso and Souloumiac, 1996; Kuleshov et al., 2015), and matrix manifold optimization methods (Afsari and Krishnaprasad, 2004; Colombo and Vlassis, 2016). Jacobi-like algorithms rely on the decomposition of the variables into single-parameter matrices (such as Givens rotations), whereas in a matrix manifold approach the objective (3) is optimized directly on the matrix manifold. As demonstrated recently (Colombo and Vlassis, 2016), a Gauss-Newton method that optimizes (3) directly on the matrix manifold $\mathbb{O}(d)$ can outperform the Jacobi-like method in terms of runtime by, roughly, one order of magnitude, for a statistically equivalent quality of the computed solutions. Finally, the problem of obtaining global guarantees for joint matrix decomposition algorithms has been considered by Kuleshov et al. (2015), but only for the case of matrix joint diagonalization. To the best of our knowledge, our work is the first that provides global solution guarantees for the joint matrix triangularization problem, corroborating the strong empirical results that have been reported in the literature (Haardt and Nossek, 1998; Abed-Meraim and Hua, 1998).

1.3 Conventions

All matrices, vectors and numbers are real. Let A be a $d \times d$ matrix, then A^T is the transpose of A , A^{-1} is the inverse of A and A^{-T} is the inverse of the transpose of A . A_{ij} (or $[A]_{ij}$) is the (i, j) entry of A . The i th singular value of A is denoted by $\sigma_i(A)$ and $\kappa(A) = \frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$ is the condition number of A . The matrix commutator $[A, B]$ is defined by $[A, B] = AB - BA$ and $\|A\|$ is the Frobenius norm defined by $\|A\|^2 = \text{Tr}(A^T A) = \sum_{i,j} A_{ij}^2$. The Kronecker product is denoted by \otimes . Depending on the context, we use $\mathbf{1}$ to denote a vector of ones or the identity matrix. $\mathbb{O}(d)$ is the manifold of orthogonal matrices U defined by $U^T U = \mathbf{1}$. $T_{\mathbb{O}(d)}$ is the tangent space of $\mathbb{O}(d)$, i.e. the set of skew-symmetric matrices satisfying $A = -A^T$. $\text{vec}(A)$ is the column wise vectorization of A . $\text{low}(A)$ and $\text{up}(A)$ are the strictly lower-diagonal and strictly upper-diagonal part of A defined by

$$[\text{low}(A)]_{ij} = \begin{cases} A_{ij} & \text{if } i > j \\ 0 & \text{if } i \leq j \end{cases} \quad (5)$$

$$[\text{up}(A)]_{ij} = \begin{cases} A_{ij} & \text{if } i < j \\ 0 & \text{if } i \geq j \end{cases} \quad (6)$$

$\text{Low} \in \{0, 1\}^{n^2 \times n^2}$ and $\text{Up} \in \{0, 1\}^{n^2 \times n^2}$ are linear operators defined by $\text{vec}(\text{low}(A)) = \text{Low} \text{vec}(A)$ and $\text{vec}(\text{up}(A)) = \text{Up} \text{vec}(A)$ respectively. $P_{\text{Low}} \in \{0, 1\}^{\frac{n(n-1)}{2} \times n^2}$ is the projector to the sub-space of (vectorized) strictly lower-diagonal matrices defined by $P_{\text{Low}} P_{\text{Low}}^T = \mathbf{1}$ and $P_{\text{Low}}^T P_{\text{Low}} = \text{Low}$. For example, letting $d = 4$, one has

$$\text{Low} = \text{diag}([0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0]) = \text{diag}(\mathbf{1}^T P_{\text{Low}}^T) \quad (7)$$

$$P_{\text{Low}} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (8)$$

and similarly for Up and P_{Up} .

2 Exact joint triangularizers

Consider the set of simultaneously diagonalizable matrices $\mathcal{M}_\circ = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$, with \hat{M}_n defined in (1). A joint triangularizer of \mathcal{M}_\circ is an orthogonal matrix U_\circ such that

$$\text{low}(U_\circ^T M_n U_\circ) = 0 \quad \text{for all } n = 1, \dots, N \quad (9)$$

The condition under which \mathcal{M}_\circ admits a finite number of joint triangularizers is established by the following lemma.

Lemma 1. *Let $\mathcal{M}_\circ = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$, with \hat{M}_n defined in (1). Then if*

$$\text{for every } i \neq i' \quad \text{there exists } n \in \{1, \dots, N\} \quad \text{s.t. } \Lambda_{ni} \neq \Lambda_{ni'} \quad (10)$$

\mathcal{M}_\circ admits $2^d d!$ exact joint triangularizers.

3 A priori perturbation analysis

Consider the approximate joint triangularization problem defined in (2) and the expansion (4). Two theoretical bounds are provided in this section. The first one is an inequality for the parameter α . The second one is an expression for the skew-symmetric matrix $X = -X^T$ that appears in (4). The explicit form of X is given in terms of the ground-truth matrices M_n and the noise matrices σW_n . Both bounds are valid up to second order terms in the perturbation parameters α and σ , i.e. they hold up to $O((\alpha + \sigma)^2)$ terms.

Theorem 1. *Let $\mathcal{M}_\sigma = \{\hat{M}_n\}_{n=1}^N$ and $\mathcal{M}_\circ = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$ with \hat{M}_n defined in (1). Assume \mathcal{M}_\circ is such that (10) is satisfied. Then there exists U_\circ , which is an exact joint triangularizer of \mathcal{M}_\circ , such that an approximate joint triangularizer of \mathcal{M}_σ can be written as*

$$U = U_\circ e^{\alpha X} \quad X = -X^T \quad \|X\| = 1 \quad (11)$$

with $\alpha > 0$ obeying

$$\alpha \leq 2\sqrt{2}\sigma \|\tilde{T}^{-1}\|_2 \sqrt{\sum_{n=1}^N \|M_n\|^2} \sqrt{\sum_{n=1}^N \|W_n\|^2} + O((\alpha + \sigma)^2) \quad (12)$$

where M_n and W_n are defined in (1), $\tilde{T} = \sum_{n=1}^N \tilde{t}_n^T \tilde{t}_n$ with $\tilde{t}_n = P_{\text{low}}(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n U_\circ \otimes 1)P_{\text{low}}^T$, and $\|\tilde{T}^{-1}\|_2$ is the spectral norm of the inverse of \tilde{T} .

It is possible to find a more explicit upper bound of (12), given in terms of the ground matrices and σ . This result is provided by the following lemma

Lemma 2. *Let α be defined as in Theorem 1, then*

$$\alpha \leq \frac{2\sigma \sqrt{d(d-1)} \kappa(V)^4}{\gamma} \sqrt{\sum_{n=1}^N \|M_n\|^2} \sqrt{\sum_{n=1}^N \|W_n\|^2} + O((\alpha + \sigma)^2) \quad \gamma = \min_{i < i'} \sum_{n=1}^N (\Lambda_{ni} - \Lambda_{ni'})^2 \quad (13)$$

where V , M_n , W_n and Λ are defined in (1).

Theorem 2. *Let $U = U_\circ e^{\alpha X}$ be the approximate joint triangularizer defined in Theorem 1. An approximate expression for the matrix αX is given by*

$$\alpha X = E - E^T \quad E = \text{mat}(P_{\text{low}}^T x) \quad x = -\sigma \left(\sum_{n=1}^N \tilde{t}_n^T \tilde{t}_n \right)^{-1} \sum_{n=1}^N \tilde{t}_n^T P_{\text{low}} \text{vec}(U_\circ^T W_n U_\circ) + O((\alpha + \sigma)^2) \quad (14)$$

where $\tilde{t}_n = P_{\text{low}}(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n U_\circ \otimes 1)P_{\text{low}}^T$, with M_n and W_n defined in (1).

Remarks on the theorems: The proof of these bounds is based on a first-order characterization of the approximate joint triangularizer U , which is defined as a stationary point of (3). The inequalities on the parameter α come from the analysis of the associated stationarity equation $\nabla \mathcal{L} = 0$, via a first order expansion around U_\circ , an exact joint triangularizer of $\mathcal{M}_{\sigma=0}$.

4 *A posteriori* perturbation analysis

The result of this section is an *a posteriori* bound on the magnitude of the approximation error:

Theorem 3. Let $\mathcal{M}_\sigma = \{\hat{M}_n\}_{n=1}^N$ and $\mathcal{M}_o = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$ with \hat{M}_n defined in (1). Assume that \mathcal{M}_o satisfies (10) and the noise matrices W_n defined in (1) obey $\|W_n\| \leq 1$. Let U be a feasible solution of the optimization problem (3). Then there exists U_o , which is an exact joint triangularizer of \mathcal{M}_o , such that U can be written as

$$U = U_o e^{\alpha X}, \quad X = -X^\top, \quad \|X\| = 1, \quad \alpha > 0, \quad (15)$$

with α obeying

$$\alpha \leq \sqrt{2} \|\beta\| \|\hat{T}_\beta^{-1}\|_2 (\sqrt{\mathcal{L}(U)} + \sigma \sqrt{N}) + O((\sigma + \alpha)^2) \quad (16)$$

where $\beta = [\beta_1, \dots, \beta_N] \in \mathbf{R}^N$, $\hat{T}_\beta = \sum_{n=1}^N \beta_n P_{\text{low}}(1 \otimes U^T \hat{M}_n^T U - U^T \hat{M}_n^T U \otimes 1) P_{\text{low}}^T$, $\|T_\beta^{-1}\|_2$ is the spectral norm of T_β^{-1} and $\mathcal{L}(U)$ is defined in (3).

Remarks on the theorem: Assuming an a priori knowledge of σ , the inequality depends only on quantities that can be computed from the observed matrices \hat{M}_n . The technique we have used to obtain the *a posteriori* bound follows an idea of Konstantinov et al. (1994) and is based on the perturbation equation

$$U^\top \left(\sum_{n=1}^N \beta_n (M_n + \sigma W_n) \right) U = \sum_{n=1}^N \beta_n (T_n + \varepsilon_n), \quad \text{low}(T_n) = 0, \quad \varepsilon_n = \text{low} \left(U^\top \left(\sum_{n=1}^N \beta_n \hat{M}_n \right) U \right) \quad (17)$$

where $\beta = [\beta_1, \dots, \beta_N]$. The difference from the single matrix case studied by Konstantinov et al. (1994) is that the lower-diagonal terms ε_n may be nonzero because an exact joint triangularizer may not exist.

5 Global guarantees for locally convergent algorithms

The existence of at least one approximate joint triangularizer of $\mathcal{M}_\sigma = \{\hat{M}_n\}_{n=1}^N$ that is close to an exact triangularizer of $\mathcal{M}_o = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$ is guaranteed by the continuity of (3) in the noise parameter σ . The distance between such an approximate joint triangularizer, U , and the exact triangularizer, U_o , is bounded by the Theorem 1. If it is possible to compute a good initialization, a locally convergent algorithm is expected to converge to such U . The following theorem provides a way to compute such a good initialization, under certain conditions on the noise parameter σ .

Theorem 4. Let $\mathcal{M}_\sigma = \{\hat{M}_n\}_{n=1}^N$ and $\mathcal{M}_o = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$ with \hat{M}_n defined in (1). Assume that \mathcal{M}_o satisfies (10) and the noise matrices W_n defined in (1) obey $\|W_n\| \leq 1$. Let $\beta = [\beta_1, \dots, \beta_N] \in \mathbf{R}^N$ such that

$$\min_{i < i'} |\text{Re}(\lambda_i(\hat{M}_\beta) - \lambda_{i'}(\hat{M}_\beta))| > 0 \quad \hat{M}_\beta = \sum_{n=1}^N \beta_n \hat{M}_n, \quad (18)$$

then a descent algorithm initialized with an orthogonal matrix U_{init} such that $\text{low}(U_{\text{init}}^T \hat{M}_\beta U_{\text{init}}) = 0$ (obtained via the Schur decomposition of \hat{M}_β) converges to an approximate joint triangularizer defined by Theorem 1 if the noise parameter σ obeys

$$\sigma \leq \frac{2\varepsilon}{\sqrt{2N} \|\hat{T}_\beta^{-1}\|_2 A_\alpha + A_\sigma} + O(\sigma^2) \quad (19)$$

where

$$\varepsilon = \frac{\gamma}{2\kappa(V)^4} \quad \gamma = \min_{i < i'} \sum_{n=1}^N (\Lambda_{ni} - \Lambda_{ni'})^2 \quad (20)$$

$$\hat{T}_\beta = P_{\text{low}}(1 \otimes U_{\text{init}}^T \hat{M}_\beta^T U_{\text{init}} - U_{\text{init}}^T \hat{M}_\beta U_{\text{init}} \otimes 1) P_{\text{low}} \quad A_\alpha = 32 \sum_{n=1}^N \|M_n\|^2 \quad A_\sigma = 16\sqrt{N} \sqrt{\sum_{n=1}^N \|M_n\|^2} \quad (21)$$

with M_n , V and Λ defined in (1).

Remarks on the theorem: The proof of the theorem consists of two steps:

(i) We first characterize the convex region containing an exact joint triangularizer U_\circ , in terms of α_{max} , the distance from U_\circ . This is obtained by requiring that the Hessian of (3) computed at $U = U_\circ e^{\alpha X}$ is positive definite for all X (with $\|X\| = 1$) if $\alpha \leq \alpha_{max}$.

(ii) Then we find a condition on the noise parameter for which the orthogonal matrix U_{init} , which is used to initialize the algorithm, belongs to the convex region characterized in the previous step. Letting $U_{init} = U_\circ e^{\alpha_{init} X_{init}}$, this is equivalent to $\alpha_{init} \leq \alpha_{max}$.

Global success guarantees for the solution U computed by a local hill-climbing algorithm can be obtained by combining Theorem 4 and Theorem 1.

6 Applications to tensor decomposition

6.1 Observable matrices

Consider an order the $N \times N \times N$ tensor of the form

$$\hat{\mathbb{T}} = \mathbb{T} + \sigma \mathbb{E} \quad \mathbb{T}_{nn'n''} = \sum_{i=1}^d Z_{ni} Z_{n'i} Z_{n''i} \quad n, n', n'' = 1, \dots, N \quad (22)$$

where $\sigma > 0$ and \mathbb{E} is an arbitrary noise term satisfying $\|\mathbb{E}\| \leq \varepsilon$, with $\|\mathbb{E}\|^2 = \sum_{nn'n''} \mathbb{E}_{nn'n''}^2$. We define the $d \times d$ ‘observable’ matrices associated with the tensor $\hat{\mathbb{T}}$ as

$$\hat{M}_n = \hat{m}_n \hat{m}_n^{-1} \quad n = 1, \dots, N \quad \hat{m} = \sum_{n=1}^N \hat{m}_n \quad (23)$$

where, for general $d \leq N$, \hat{m}_n are dimension-reduced tensor slices defined by

$$\hat{m}_n = U_d^T \tilde{m}_n V_d \quad [\tilde{m}_n]_{n'n''} = \hat{\mathbb{T}}_{nn'n''} \quad n, n', n'' = 1, \dots, N \quad (24)$$

with U_d and V_d being $N \times d$ Stiefel matrices obtained by staking the first d left and right singular vectors of $\sum_{n=1}^N \tilde{m}_n$. The definition (23) makes sense only if \hat{m} is invertible, *i.e.* if the d th singular value of $\sum_{n=1}^N \tilde{m}_n$ is non-vanishing. Assuming $d = N$ there is no need of introducing the dimension reduction matrices U_d and V_d and the observable matrices are then defined by

$$\hat{M}_n = \hat{m}_n \hat{m}_n^{-1} \quad [\hat{m}_n]_{n'n''} = \hat{\mathbb{T}}_{nn'n''} \quad n, n', n'' = 1, \dots, N \quad (d = N) \quad \hat{m} = \sum_{n=1}^N \hat{m}_n \quad (25)$$

where \hat{m} is assumed to be invertible. A more general definition of \hat{m} would be $\hat{m}_\theta = \sum_{n=1}^N \theta_n \hat{m}_n$ where θ is an arbitrary N -dimensional vector. In what follows we consider the case $d = N$ and $\theta = 1$ but generalizations to $d \leq N$ and $\theta \neq 1$ are straightforward. Observable matrices of the form (23) cannot be defined if $d > N$. Given (22) and (25), it is easy to prove the following lemma

Lemma 3. *If Z is invertible and $[1^T Z]_i \neq 0$ for all $i = 1, \dots, d$, the observable matrices \hat{M}_n defined in (25) can be expanded as follows*

$$\hat{M}_n = M_n + \sigma W_n + O(\sigma^2) \quad M_n = Z \text{diag}(\mathbf{e}_n^T Z) (\text{diag}(1^T Z))^{-1} Z^{-1} \quad W_n = e_n m^{-1} + m_n m^{-1} e m^{-1} \quad (26)$$

where $n = 1, \dots, N$, the vector \mathbf{e}_n is the n th basis vector, and

$$[e_n]_{n'n''} = E_{nn'n''} \quad e = \sum_{n=1}^N e_n \quad [m_n]_{n'n''} = \mathbb{T}_{nn'n''} \quad m = \sum_{n=1}^N m_n. \quad (27)$$

If \mathbb{E} in (22) obeys $\|\mathbb{E}\| \leq \varepsilon$, then

$$\|M_n\| \leq \frac{d \kappa(Z)^2 \max |Z|}{\min |1^T Z|} \quad \|W_n\| \leq \frac{\varepsilon \kappa(Z)^2 \sqrt{d}}{\|Z\|^2 \min |1^T Z|} \left(1 + \frac{d \kappa(Z)^2 \max |Z|}{\min |1^T Z|} \right) \quad (28)$$

6.2 Estimation of the tensor components Z

Lemma 3 implies that Z can be obtained, up to normalization constants, from the estimated joint eigenvalues of the nearly jointly diagonalizable matrices (26). Let U be an approximate joint triangularizer of $\mathcal{M}_\sigma = \{\hat{M}_n\}_{n=1}^N$ obeying the bound in Theorem 1. The corresponding estimation of Z is given by

$$\frac{Z_{ni}^*}{[1^T Z^*]_i} = [U^T \hat{M}_n U]_{ii} \quad n = 1, \dots, N \quad i = 1, \dots, d \quad (29)$$

where $[1^T Z^*]_i$ is an undetermined column-rescaling factor and we assume $N = d$. Under the conditions that Z is invertible and $[1^T Z]_i \neq 0$ for all $i = 1, \dots, d$, the difference between the estimated tensor components (29) and the ground-truth tensor components Z is bounded by the following theorem.

Theorem 5. *Let $\hat{\mathbb{T}}$ be the tensor defined in (22) and assume $N = d$, Z is invertible, and $[1^T Z]_i \neq 0$ for all $i = 1, \dots, d$. Let U be an approximate joint triangularizer of $\mathcal{M}_\sigma = \{\hat{M}_n\}_{n=1}^N$, with \hat{M}_n defined in (26), and $\frac{Z_{ni}^*}{[1^T Z^*]_i} = [U^T \hat{M}_n U]_{ii}$ for all $n = 1, \dots, N$ and $i = 1, \dots, d$. Then, if U obeys the bound in Theorem 1, Z_* is such that*

$$\left| \frac{Z_{ni}^*}{[1^T Z^*]_i} - \frac{Z_{ni}}{[1^T Z]_i} \right| \leq 4\sigma \frac{\sqrt{d(d-d)}\kappa(Z)^4}{\gamma} M^2 W + \sigma W + O(\sigma^2) \quad (30)$$

where

$$\gamma = \frac{1}{N} \min_{i \neq i'} \sum_{n=1}^N (Z_{ni} - Z_{ni'})^2 \quad M \leq \frac{N\kappa(Z)^2 \max |Z|}{\min |1^T Z|} \quad W \leq \frac{\varepsilon \sqrt{N} \kappa(Z)^2}{\|Z\|^2 \min |1^T Z|} \left(1 + \frac{N\kappa(Z)^2 \max |Z|}{\min |1^T Z|} \right) \quad (31)$$

Remark on the theorem: Theorem 5 provides a first order approximation of the estimation error and it is valid up to terms proportional to σ^2 . The assumption on $[1^T Z]_i$ can be relaxed by defining \hat{m} as $\hat{m}_\theta = \sum_{n=1}^N \theta_n \hat{m}_n$, where θ is any N -dimensional vector for which $[\theta^T Z]_i \neq 0$ for all $i = 1, \dots, d$. The normalization constants $[\theta^T Z^*]_i$ can then be obtained from $\hat{m}_\theta = \sum_{n=1}^N \theta_n \hat{m}_n$ and the corresponding estimates $\frac{Z_{ni}^*}{[\theta^T Z^*]_i}$ by solving the following matrix equation

$$\hat{m}_\theta = Z^* \frac{1}{\text{diag}(\theta^T Z^*)} \text{diag}(\theta^T Z^*)^3 \left(Z^* \frac{1}{\text{diag}(\theta^T Z^*)} \right)^T \left[Z^* \frac{1}{\text{diag}(\theta^T Z^*)} \right]_{ni} = [U^T \hat{M}_n U]_{ii} \quad (32)$$

Finally, by using (29) and the *a posteriori* error analysis of Section 4 it is possible to obtain analogous bounds that depend only on the observable matrices (25).

7 Other lemmas and proofs

7.1 Proof of Lemma 1

Lemma 1 establishes a sufficient condition for the existence of $2^d d!$ exact joint triangularizers of $\mathcal{M}_\circ = \{\hat{M}_n |_{\sigma=0}\}_{n=1}^N$, with \hat{M}_n defined in (1). The proof consists of showing that, if (10) holds (i) there exist $2^d d!$ exact joint triangularizers of \mathcal{M}_\circ and (ii) it is impossible to find more than $2^d d!$ such orthogonal matrices. Lemma 4 can be used to prove that, when (10) is fulfilled, it is possible to define a linear combination of the matrices $M_n \in \mathcal{M}_\circ$ with distinct eigenvalues. Let M be such linear combination of the matrices M_n . Since any real $d \times d$ matrix with distinct eigenvalues admits $2^d d!$ triangularizers, M admits $2^d d!$ triangularizers. Now, since $[M_n, M_{n'}] = 0$ one has

$$[M, M_n] = 0 \quad \forall n = 1, \dots, N \quad (33)$$

implying that all $2^d d!$ triangularizers of M exactly triangularize all $M_n \in \mathcal{M}_\circ$. This is due to the fact that commuting matrices are always joint triangularizable and implies that \mathcal{M}_\circ has at least $2^d d!$ joint triangularizers. But the commutation relation (33) also implies that any possible additional triangularizer of a matrix $M_n \in \mathcal{M}_\circ$ would exactly triangularize M . This contradicts the fact that M admits only $2^d d!$ exact triangularizers and proves the lemma. \square

7.2 Proof of Theorem 1

The stationary point of (3) are defined by the equation $\nabla \mathcal{L} = 0$ where $\nabla \mathcal{L}$ is the gradient of \mathcal{L} and \mathcal{L} is defined in (3). According to Lemma 5, if U is a stationary point of (3), then

$$\nabla \mathcal{L} = S - S^T = 0 \quad S = \sum_{n=1}^N \left[U^T \hat{M}_n^T U, \text{low}(U^T \hat{M}_n U) \right] \quad (34)$$

Now, let $U = U_\circ e^{\alpha X}$, where U_\circ is an exact triangularizer of $\mathcal{M}_\circ = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$, \hat{M}_n are defined in (1), $X = -X^T$ and one can assume $\|X\| = 1$ and $\alpha > 0$. The expansion of S in α and σ reads

$$S = S|_{(\alpha=0, \sigma=0)} + \alpha \partial_\alpha S|_{\sigma=0} + \sigma \partial_\sigma S|_{\alpha=0} + O((\alpha + \sigma)^2) \quad (35)$$

$$= \sum_{n=1}^N [U_\circ^T M_n^T U_\circ, \text{low}([U_\circ^T M_n U_\circ, \alpha X])] + \sum_{n=1}^N [U_\circ^T M_n^T U_\circ, \text{low}(U_\circ^T \sigma W_n U_\circ)] + O((\alpha + \sigma)^2) \quad (36)$$

where we have defined $\partial_\alpha f = \frac{\partial}{\partial \alpha} f|_{\alpha=0}$ and $\partial_\sigma f = \frac{\partial}{\partial \sigma} f|_{\sigma=0}$. Note that, for all $n = 1, \dots, N$, $[U_\circ^T M_n^T U_\circ, \text{low}(A)]$ is strictly lower-triangular for any A because $\text{up}(U_\circ^T M_n^T U_\circ) = 0$. The latter follows from the fact that U_\circ is an exact triangularizer of \mathcal{M}_\circ and hence $U_\circ^T M_n U_\circ$ is upper triangular, for all $n = 1, \dots, N$. Considering only the lower-diagonal part of the stationarity equation one obtains the necessary condition

$$0 = \text{low}(S - S^T) = \text{low}(\alpha \partial_\alpha S|_{\sigma=0} + \sigma \partial_\sigma S|_{\alpha=0}) + O((\alpha + \sigma)^2) \quad (37)$$

since the first order terms of S^T are upper triangular. The projected stationarity equation (37) reads

$$\text{low} \left(\sum_{n=1}^N [U_\circ^T M_n^T U_\circ, \text{low}([U_\circ^T M_n U_\circ, \alpha X])] \right) = -\text{low} \left(\sum_{n=1}^N [U_\circ^T M_n^T U_\circ, \text{low}(U_\circ^T W_n U_\circ)] \right) + O((\alpha + \sigma)^2) \quad (38)$$

Moreover, since $\text{low}(U_\circ^T M_n U_\circ) = 0$ for all $n = 1, \dots, N$ one has

$$\text{low}([U_\circ^T M_n U_\circ, \alpha X]) = \text{low}([U_\circ^T M_n U_\circ, \text{low}(\alpha X)]) \quad (39)$$

This means that the linear operator defined by

$$\mathcal{S} \text{low}(X) = \text{low} \left(\sum_{n=1}^N [U_\circ^T M_n^T U_\circ, \text{low}([U_\circ^T M_n U_\circ, \text{low}(X)])] \right) \quad (40)$$

maps the subspace of strictly lower dimensional matrices into itself. This is a $\frac{d(d-1)}{2}$ -dimensional subspace that has the same degrees of freedom as the set of $d \times d$ skew-symmetric matrices. Each $d \times d$ skew-symmetric matrix is mapped into this subspace by means of the projection $P_{\text{low}} \text{vec}(X)$. Conversely, letting x be an element of this subspace, the corresponding $d \times d$ skew-symmetric matrix X is given by $X = \text{mat}(P_{\text{low}}^T x) - \text{mat}(P_{\text{low}} x)^T$. Let T be the linear operator defined by the vectorization of (38)

$$T = \sum_{n=1}^N t_n^T t_n \quad t_n = \text{Low}(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n U_\circ \otimes) \text{Low} \quad (41)$$

Its reduction to the subspace of strictly lower-diagonal matrices can be written as

$$\tilde{T} = P_{\text{low}} T P_{\text{low}}^T = \sum_{n=1}^N \tilde{t}_n^T \tilde{t}_n \quad \tilde{t}_n = P_{\text{low}}(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n U_\circ \otimes) P_{\text{low}}^T \quad (42)$$

Then one has

$$P_{\text{low}} \text{vec}(\mathcal{S} \text{low}(\alpha X)) = \tilde{T} P_{\text{low}} \text{vec}(\alpha X) \quad (43)$$

The $\frac{d(d-1)}{2} \times \frac{d(d-1)}{2}$ \tilde{T} is positive definite if the non-degeneracy condition in (10) is fulfilled (see Lemma 6). Under this assumption

$$\alpha P_{\text{low}} \text{vec}(X) = -\tilde{T}^{-1} P_{\text{low}} \text{vec} \left(\sum_{n=1}^N [U_\circ^T M_n^T U_\circ, \text{low}(U_\circ^T W_n U_\circ)] \right) + O((\alpha + \sigma)^2) \quad (44)$$

Taking the norm of both sides one has

$$\alpha \leq 2\sqrt{2}\sigma \|\tilde{T}^{-1}\|_2 \sqrt{\sum_{n=1}^N \|M_n\|^2} \sqrt{\sum_{n=1}^N \|W_n\|^2} \quad (45)$$

where we have used $\|\text{low}(X)\| = \frac{1}{\sqrt{2}}\|X\|$, $\|X\| = 1$ and

$$\left\| P_{\text{low}} \text{vec} \left(\sum_{n=1}^N [U_\circ^T M_n^T U_\circ, \text{low}(U_\circ^T W_n U_\circ)] \right) \right\| \leq 2\sigma \sqrt{\sum_{n=1}^N \|M_n\|^2} \sqrt{\sum_{n=1}^N \|W_n\|^2} \quad (46)$$

from $\|\sum_{n=1}^N t_n^T \text{vec}(U_\circ^T \sigma W_n U_\circ)\| \leq \sqrt{\sum_{n=1}^N \|t_n^T\|^2} \sqrt{\sum_{n=1}^N \|\sigma W_n\|^2}$, $\|t_n\|^2 \leq 4\|M_n\|^2$. \square

7.3 Proof of Lemma 2

Consider the inequality on the perturbation parameter α given in (45). Lemma 6 states that the matrix \tilde{T} is positive definite if the non-degeneracy condition (10) is fulfilled and in this case

$$\|\tilde{T}^{-1}\|_2 \leq \sqrt{\frac{d(d-1)}{2}} \frac{\kappa(V)^4}{\gamma} \quad \gamma = \min_{i < i'} \sum_{n=1}^N (\Lambda_{ni} - \Lambda_{ni'})^2 \quad (47)$$

This implies

$$\alpha \leq \frac{2\sigma\sqrt{d(d-1)}\kappa(V)^4}{\gamma} \sqrt{\sum_{n=1}^N \|M_n\|^2} \sqrt{\sum_{n=1}^N \|W_n\|^2} \quad (48)$$

□

7.4 Proof of Theorem 2

Theorem 2 follows from (44) where one can use

$$P_{\text{low}} \text{vec} \left(\sum_{n=1}^N [U_{\circ}^T M_n^T U_{\circ}, \text{low}(U_{\circ}^T \sigma W_n U_{\circ})] \right) = \sigma \sum_{n=1}^N \tilde{t}_n^T P_{\text{low}}(U_{\circ}^T W_n U_{\circ}) \quad (49)$$

to obtain

$$P_{\text{low}} \text{vec}(\alpha X) = -\sigma \left(\sum_{n=1}^N \tilde{t}_n^T \tilde{t}_n \right)^{-1} \sum_{n=1}^N \tilde{t}_n^T P_{\text{low}}(U_{\circ}^T W_n U_{\circ}) \quad (50)$$

with $\tilde{t}_n = P_{\text{low}}(1 \otimes U_{\circ}^T M_n^T U_{\circ} - U_{\circ}^T M_n U_{\circ} \otimes 1) P_{\text{low}}^T$. □

7.5 Proof of Theorem 3

Let $\sum_{n=1}^N \beta_n \hat{M}_n$ be a general linear combination of the input matrices, where β_n , $n = 1, \dots, N$ are arbitrary real numbers. Let U_{\circ} be an exact joint triangularizer of \mathcal{M}_{\circ} , and U be a feasible solution of the joint triangularization problem (3). By construction U is an orthogonal matrix and can be written as $U = U_{\circ} e^{\alpha X}$, with $X = -X^T$, $\|X\| = 1$ and $\alpha > 0$. For any choice of β one has

$$U^T \left(\sum_{r=1}^N \beta_r \hat{M}_r \right) U = \sum_{r=1}^N \beta_r (\hat{T}_r + \varepsilon_r) \quad \text{low}(\hat{T}_r) = 0 \quad \varepsilon_r = \text{low}(U^T \hat{M}_r U) \quad (51)$$

By projecting onto the strictly lower-diagonal part and considering the expansion $U = U_{\circ} e^{\alpha X}$, we obtain

$$\sum_{r=1}^N \beta_r \varepsilon_r = \sum_{r=1}^N \beta_r \text{low} \left(e^{-\alpha X} U_{\circ}^T M_r U_{\circ} e^{\alpha X} + e^{-\alpha X} U_{\circ}^T \sigma W_r U_{\circ} e^{\alpha X} \right) \quad (52)$$

$$= \sum_{r=1}^N \beta_r \text{low} \left([U_{\circ}^T M_r U_{\circ}, \alpha X] + U_{\circ}^T \sigma W_r U_{\circ} \right) + O((\alpha + \sigma)^2) \quad (53)$$

For any X , one has $\text{low}([U_{\circ}^T M_n U_{\circ}, X]) = \text{low}([U_{\circ}^T M_n U_{\circ}, \text{low}(X)])$ because $U_{\circ}^T M_n U_{\circ}$ is upper triangular. The identity (53) can be rewritten as

$$\text{low} \left(\left[U_{\circ}^T \sum_{r=1}^N \beta_r M_r U_{\circ}, \text{low}(\alpha X) \right] \right) = \sum_{r=1}^N \beta_r \text{low}(\varepsilon_r - U_{\circ}^T \sigma W_r U_{\circ}) \quad (54)$$

whose vectorization reads

$$T_{\beta} \text{vec}(\alpha X) = \text{vec} \left(\text{low} \left(\sum_{r=1}^N \beta_r \varepsilon_r - \sigma W_{\beta} \right) \right) \quad T_{\beta} = \text{Low}(1 \otimes M_{\beta}^T - M_{\beta} \otimes 1) \text{Low} \quad (55)$$

where $M_{\beta} = \sum_{r=1}^N \beta_r U_{\circ}^T M_r U_{\circ}$ and $W_{\beta} = \sum_{r=1}^N \beta_r U_{\circ}^T W_r U_{\circ}$. The reduction of T_{β} to the subspace of strictly lower-diagonal matrices is

$$\tilde{T}_{\beta} = P_{\text{low}} T_{\beta} P_{\text{low}}^T = P_{\text{low}}(1 \otimes M_{\beta}^T - M_{\beta} \otimes 1) P_{\text{low}}^T \quad (56)$$

Lemma 8 can be used to show that \tilde{T}_β is invertible if M_β is invertible and $\lambda_i(M_\beta) \neq \lambda_{i'}(M_\beta)$ for all $i \neq i'$. Under this assumption one can write

$$\alpha \text{vec}(X) = \tilde{T}_\beta^{-1} \text{vec} \left(\text{low} \left(\sum_{r=1}^N \beta_n \varepsilon_n - \sigma W_\beta \right) \right) + O((\alpha + \sigma)^2) \quad (57)$$

and, by taking the norm in both sides,

$$\alpha \leq \sqrt{2} \|\tilde{T}_\beta^{-1}\|_2 \|\beta\| \left(\sqrt{\sum_{r=1}^N \|\varepsilon_n\|^2} + \sigma \sqrt{\sum_{n=1}^N \|W_n\|^2} \right) + O((\alpha + \sigma)^2) \quad (58)$$

$$\leq \sqrt{2} \|\tilde{T}_\beta^{-1}\|_2 \left(\sqrt{\mathcal{L}(U)} + \sigma \sqrt{N} \right) + O((\alpha + \sigma)^2) \quad (59)$$

where we have used the assumption $\|\beta\| = 1$ and $\|W_n\| \leq 1$. Finally, one has

$$\tilde{T}_\beta = \sum_{r=1}^N \beta_n P_{\text{low}} (1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n^T U_\circ \otimes 1) P_{\text{low}}^T \quad (60)$$

$$= \sum_{r=1}^N \beta_n P_{\text{low}} (1 \otimes U_\circ^T \hat{M}_n^T U_\circ - U_\circ^T \hat{M}_n^T U_\circ \otimes 1) P_{\text{low}}^T + O(\sigma) \quad (61)$$

$$= \sum_{r=1}^N \beta_n P_{\text{low}} (1 \otimes U^T \hat{M}_n^T U - U^T \hat{M}_n^T U \otimes 1) P_{\text{low}} + O(\sigma + \alpha) \quad (62)$$

$$= \hat{T}_\beta + O(\sigma + \alpha) \quad (63)$$

where we have defined $\hat{T}_\beta = \sum_{r=1}^N \beta_n P_{\text{low}} (1 \otimes U^T \hat{M}_n^T U - U^T \hat{M}_n^T U \otimes 1) P_{\text{low}}$. It follows that $\|\tilde{T}_\beta^{-1}\|_2 = \|\hat{T}_\beta^{-1}\|_2 + O(\sigma + \alpha)$ and hence

$$\alpha \leq \sqrt{2} \|\hat{T}_\beta^{-1}\|_2 \left(\sqrt{\mathcal{L}(U)} + \sigma \sqrt{N} \right) + O((\alpha + \sigma)^2) \quad (64)$$

□

7.6 Proof of Theorem 4

The Hessian of (3) at U is positive definite if, for all X such that $X = -X^T$, $\langle X, \nabla^2 \mathcal{L} X \rangle > 0$, where

$$\langle X, \nabla^2 \mathcal{L}(U) X \rangle = \frac{d^2}{dt^2} \mathcal{L}(U e^{tX})|_{t=0} \quad (65)$$

Lemma 9 shows that this is the case if

$$U = U_\circ e^{\alpha Y} \quad Y = -Y^T \quad \|Y\| = 1 \quad \alpha \leq \alpha_{\max} \quad \alpha_{\max} = \frac{2\varepsilon - \sigma A_\sigma}{A_\alpha} + O((\alpha + \sigma)^2) \quad (66)$$

$$\varepsilon = \frac{\gamma}{2\kappa(V)^4} \quad \gamma = \min_{j < j'} \sum_{n=1}^N (\Lambda_{nj} - \Lambda_{nj'})^2 \quad A_\alpha = 32 \sum_{n=1}^N \|M_n\|^2 \quad A_\sigma = 16\sqrt{N} \sqrt{\sum_{n=1}^N \|M_n\|^2} \quad (67)$$

where we have assumed $\|W_n\| \leq 1$. The condition under which the Hessian of (3) at U_\circ is positive definite is $\alpha_{\max} > 0$. If U is a minimizer of $\mathcal{L}(U)$, this condition ensures that U_\circ belongs to the convex region centered in U . Now, assume that it is possible to find a vector $\beta = [\beta_1, \dots, \beta_N]$ such that $\|\beta\| = 1$ and the operator T_β defined by

$$T_\beta = P_{\text{low}} (1 \otimes U_{\text{init}}^T \hat{M}_\beta^T U_{\text{init}} - U_{\text{init}}^T \hat{M}_\beta^T U_{\text{init}} \otimes 1) P_{\text{low}} \quad U_{\text{int}} \in \mathbb{O}(d) \text{ s.t. } \text{low}(U_{\text{int}}^T \hat{M}_\beta U_{\text{int}}) = 0 \quad \hat{M}_\beta = \sum_{n=1}^N \beta_n \hat{M}_n \quad (68)$$

is invertible. The orthogonal matrix U_{int} is defined by the Schur decomposition of \hat{M}_β . According to Lemma 8, T_β is invertible if \hat{M}_β is invertible and has real separated eigenvalues, *i.e.* if $\lambda_i(\hat{M}_\beta)$ are real for all $i = 1, \dots, d$ and $\min_{i < i'} |\lambda_i(\hat{M}_\beta) - \lambda_{i'}(\hat{M}_\beta)| > 0$. Finding such a \hat{M}_β is possible if σ is small enough. This is a consequence of Lemma 4 and standard eigenvalues perturbation results. Otherwise, the separation of the eigenvalues of \hat{M}_β can be checked numerically, since \hat{M}_β is an observable quantity. Now, let $M_\beta = \sum_{n=1}^N \beta_n M_n$, $W_\beta = \sum_{n=1}^N \beta_n W_n$ and $U_\circ \in \mathbb{O}(d)$ be such that $\text{low}(U_\circ^T M_\beta U_\circ) = 0$. By writing $U_\circ = U_{\text{init}} e^{\alpha Y}$ one has

$$U_\circ^T M_\beta U_\circ = e^{-\alpha Y} U_{\text{init}}^T (\hat{M}_\beta - \sigma W_\beta) U_{\text{init}} e^{\alpha Y} \quad (69)$$

Since $\text{low}(U_o^T M_\beta U_o) = 0$ this implies

$$\text{low}\left(e^{-\alpha Y} U_{init}^T (\hat{M}_\beta - \sigma W_\beta) U_{init} e^{-\alpha Y}\right) = 0 \quad \Rightarrow \quad \text{low}\left([U_{init}^T \hat{M}_\beta U_{init}, \alpha Y]\right) = \text{low}\left(U_{init}^T \sigma W_\beta U_{init}\right) + O(\alpha^2) \quad (70)$$

The strictly lower-diagonal part of $[A, \alpha Y]$ is equal to the strictly lower diagonal part of $[A, \text{low}(\alpha Y)]$, if A is upper-triangular. Then, by considering the projection to the subspace of strictly lower diagonal matrices of (70) (see proof of Theorem 1 for more details), one obtains

$$T_\beta P_{\text{low}} \text{vec}(\alpha Y) = P_{\text{low}} \text{vec}\left(U_{init}^T \sigma W_\beta U_{init}\right) + O(\alpha^2) \quad (71)$$

with T_β defined in (68). Since T_β is invertible one has

$$P_{\text{low}} \text{vec}(\alpha Y) = T_\beta^{-1} P_{\text{low}} \text{vec}\left(U_{init}^T \sigma W_\beta U_{init}\right) \quad (72)$$

and taking the norm in both sides

$$\alpha \leq \sqrt{2} \|T_\beta^{-1}\| \|P_{\text{low}} \text{vec}\left(U_{init}^T \sigma \hat{W}_\beta U_{init}\right)\| + O(\alpha^2) \quad (73)$$

where $\|T_\beta^{-1}\|_2$ is the spectral norm of T_β^{-1} . This implies that the initialization matrix U_{init} obtained from the Schur decomposition of \hat{M}_β can be written as $U_{init} = U_o e^{-\alpha Y}$, with α obeying

$$\alpha \leq \alpha_{init} \quad \alpha_{init} = \sigma \sqrt{2N} \|T_\beta^{-1}\|_2 + O(\alpha_{init}^2) \quad (74)$$

where we have used $\|\text{Lowvec}(U_{init}^T W_\beta U_{init})\| \leq \sqrt{N} \|\beta\| = \sqrt{N}$, since $\|W_n\| \leq 1$ and $\|\beta\| = 1$ by assumption. Now, the initialization matrix U_{init} belongs to the convex region containing U_o if $\alpha < \alpha_{max}$, with α_{max} given in (66). It follows that a descent algorithm initialized with U_{init} converges to the minimum of the convex region containing U_o if

$$\sigma \sqrt{2N} \|T_\beta^{-1}\|_2 \leq \frac{2\varepsilon - \sigma A_\sigma}{A_\alpha} + O(\sigma^2) \quad (75)$$

or equivalently

$$\sigma \leq \frac{2\varepsilon}{\sqrt{2N} \|T_\beta^{-1}\|_2 A_\alpha + A_\sigma} + O(\sigma^2) \quad (76)$$

□

7.7 Proof of Lemma 3

Let m_n be defined by $[m_n]_{n'n''} = \mathbb{T}_{nn'n''}$ for all $n, n', n'' = 1, \dots, N$. From the definition of tensor slice $[\hat{m}_n]_{n'n''} = \hat{\mathbb{T}}_{nn'n''}$ one has $\hat{m}_n = m_n + \sigma e_n$, where the noise term is defined by $[e_n]_{n'n''} = E_{nn'n''}$. Let $m = \sum_n m_n$ and $e = \sum_n e_n$, then, from the definition of \mathbb{T} given in (22) one has

$$\hat{m}_n = m_n + \sigma e_n \quad m_n = Z \text{diag}(\mathbf{e}_n^T Z) Z^T \quad \hat{m} = m + \sigma e \quad m = Z \text{diag}(1^T Z) Z^T \quad (77)$$

and

$$\hat{M}_n = \hat{m}_n \hat{m}^{-1} = (m_n + \sigma e_n)(m + \sigma e)^{-1} = m_n m^{-1} + \sigma (e_n m^{-1} + m_n m^{-1} e m^{-1}) + O(\sigma^2) \quad (78)$$

where it is easy to check that

$$m_n m^{-1} = Z \text{diag}(\mathbf{e}_n^T Z) Z^T (Z \text{diag}(1^T Z) Z^T)^{-1} = Z \text{diag}(\mathbf{e}_n^T Z) (\text{diag}(1^T Z))^{-1} Z^{-1} \quad (79)$$

where we have assumed $d = N$ and the matrices Z to be invertible. From the definitions above it follows

$$\|m_n\| = \|Z \text{diag}(\mathbf{e}_n^T Z) Z^T\| \quad (80)$$

$$\leq \|Z\|^2 \|\text{diag}(\mathbf{e}_n^T Z)\| \quad (81)$$

$$\leq \|Z\|^2 \sqrt{N} \max_n |Z_{ni}| \quad (82)$$

$$\leq \|Z\|^2 \sqrt{N} \max |Z| \quad (83)$$

and, assuming $[1^T Z]_i \neq 0$ for all $i = 1, \dots, d$,

$$\|m^{-1}\| = \left\| \left(\sum_{n=1}^N m_n \right)^{-1} \right\| \quad (84)$$

$$= \|Z^{-T} (\text{diag}(1^T Z))^{-1} Z^{-1}\| \quad (85)$$

$$\leq \|Z^{-1}\|^2 \|(\text{diag}(1^T Z))^{-1}\| \quad (86)$$

$$= \|Z^{-1}\|^2 \frac{\sqrt{N}}{\min |1^T Z|} \quad (87)$$

This implies, for all $n = 1, \dots, N$,

$$\|M_n\| = \|m_n m^{-1}\| \quad (88)$$

$$\leq \|m_n\| \|m^{-1}\| \quad (89)$$

$$\leq N\kappa(Z)^2 \frac{\max |Z|}{\min |1^T Z|} \quad (90)$$

and

$$\|W_n\| = \|e_n m^{-1} + m_n m^{-1} e m^{-1}\| \quad (91)$$

$$\leq \varepsilon \|m^{-1}\| (1 + \|m_n\| \|m^{-1}\|) \quad (92)$$

$$\leq \varepsilon \|Z^{-1}\|^2 \frac{\sqrt{N}}{\min |1^T Z|} \left(1 + N\kappa(Z)^2 \frac{\max |Z|}{\min |1^T Z|} \right) \quad (93)$$

$$\leq \varepsilon \frac{\kappa(Z)^2}{\|Z\|^2} \frac{\sqrt{N}}{\min |1^T Z|} \left(1 + N\kappa(Z)^2 \frac{\max |Z|}{\min |1^T Z|} \right) \quad (94)$$

□

7.8 Proof of Theorem 5

Lemma 3 shows that the matrices \hat{M}_n are approximately jointly diagonalizable. Let $\mathcal{M}_\sigma = \{\hat{M}_n\}_{n=1}^N$ and $\mathcal{M}_o = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$. Assume that \mathcal{M}_o is such that (10) is satisfied. In this case the solutions of (3) are characterized by the Theorem 1. Now, let U_* be a minimizer of (3), then U_* can be written as $U_* = U_o e^{\alpha_* X_*}$, with $\|X_*\| = 1$, $X_* = -X_*^T$ and α_* obeying the bound given by Theorem 1. According to (29), the approximate joint triangularizer U_* can be used to estimate the element of the tensor component Z . The distance between the estimated joint eigenvalues and the exact eigenvalues of a set of nearly jointly diagonalizable matrices is bounded by Lemma 12. Using the result of Theorem 2 and Lemma 12 with the definition (29) one obtains

$$\left| \frac{Z_{ni}^*}{[1^T Z^*]_i} - \frac{Z_{ni}}{[1^T Z]_i} \right| \leq 4N\sigma \frac{\sqrt{d(d-1)}\kappa(V)^4}{\gamma} M^2 W + \sigma W + O(\sigma^2) \quad (95)$$

for all $i = 1, \dots, d$ and all $n = 1, \dots, N$. From Lemma 3 on has

$$M \leq N\kappa(Z)^2 \frac{\max |Z|}{\min |1^T Z|} \quad W \leq \varepsilon \frac{\kappa(Z)^2}{\|Z\|^2} \frac{\sqrt{N}}{\min |1^T Z|} \left(1 + N\kappa(Z)^2 \frac{\max |Z|}{\min |1^T Z|} \right) \quad (96)$$

from which the claim of the theorem. □

7.9 Auxiliary lemmas

Lemma 4. *If (10) holds it is possible to find $\beta = [\beta_1, \dots, \beta_N]$ such that*

$$M = \sum_{n=1}^N \beta_n M_n \quad (97)$$

has real distinct eigenvalues.

Proof of Lemma 4 Let $\beta = [\beta_1, \dots, \beta_N]$, then the eigenvalues of $M = \sum_{n=1}^N \beta_n M_n$ are

$$\lambda_i(M) = \sum_{n=1}^N \beta_n \Lambda_{in} \quad i = 1, \dots, d \quad (98)$$

We want to show that (10) implies that it is possible to find β_1, \dots, β_N such that $\lambda_i \neq \lambda_{i'}$ for all $i \neq i'$, with $i, i' = 1, \dots, d$. This can be seen as follows. It is always possible to choose \tilde{m}_2 such that $\lambda_1(\tilde{m}_2) \neq \lambda_2(\tilde{m}_2)$. Now, assume that \tilde{m}_n is such that $\lambda_i(\tilde{m}_n) \neq \lambda_j(\tilde{m}_n)$ for all $i \neq j$ and $i, j \leq n$. Consider $\lambda_{n+1}(\tilde{m}_n)$. We want to show that it is possible to find a matrix m_{n+1} and a coefficient β_{n+1} such that the first $n+1$ eigenvalues of $\tilde{m}_{n+1} = \tilde{m}_n + \beta_{n+1} m_{n+1}$ are distinct, that is $\lambda_i(\tilde{m}_{n+1}) \neq \lambda_j(\tilde{m}_{n+1})$ for all $i \neq j$ and $i, j \leq n+1$. If $\lambda_{n+1}(\tilde{m}_n) \neq \lambda_i(\tilde{m}_n)$ for all $i \leq n$, one has $\tilde{m}_{n+1} = \tilde{m}_n$. Otherwise, there exists an $i \leq n$ such that $\lambda_{n+1}(\tilde{m}_n) = \lambda_i(\tilde{m}_n)$. Note that, since $\lambda_i(\tilde{m}_n) \neq \lambda_j(\tilde{m}_n)$ for all $i \neq j$ and $i, j \leq n$, there is only one such i . Let m_{n+1} be the matrix in \mathcal{M}_o satisfying $\lambda_{n+1}(\tilde{m}_n) \neq \lambda_i(\tilde{m}_n)$ and

$$\beta_{n+1} \in \mathbf{R} \quad \text{s.t.} \quad \beta_{n+1} \neq 0 \quad \text{and} \quad \beta_{n+1} \neq \frac{\lambda_i(\tilde{m}_n) - \lambda_j(\tilde{m}_n)}{\lambda_j(m_{n+1}) - \lambda_i(m_{n+1})} \quad \text{for all } i \neq j \quad i, j \leq n \quad (99)$$

Then it is easy to check that the first $n + 1$ eigenvalues of $\tilde{m}_{n+1} = \tilde{m}_n + \beta_{n+1}m_{n+1}$ are distinct. The matrix M is then constructed by repeating the above procedure until $n + 1 = d$. \square

Lemma 5. *Let U be a stationary point of (3), then*

$$S - S^T = 0 \quad S = \sum_{n=1}^N \left[U^T \hat{M}_n^T U, \text{low}(U^T \hat{M}_n U) \right] \quad (100)$$

Proof of Lemma 5 Let $f(U)$ be a function defined on $\mathbb{O}(d)$. The directional derivatives of f at U in the direction X are defined as

$$D_X f(U) = \langle \nabla f(U), X \rangle \quad (101)$$

$$= \frac{d}{dt} f(Ue^{Xt})|_{t=0} \quad (102)$$

where $X = -X^T$ and the scalar product in the tangent space is defined by $\langle A, B \rangle = \text{Tr}(A^T B)$. In particular, for (3) one has

$$\langle X, \nabla \mathcal{L}(U, \mathcal{M}_\sigma) \rangle = \frac{d}{dt} \mathcal{L}(Ue^{Xt}, \mathcal{M}_\sigma) \Big|_{t=0} \quad (103)$$

$$= \sum_{n=1}^N \text{Tr} \left([U^T \hat{M}_n^T U, X] \text{low}(U^T \hat{M}_n U) + [U^T \hat{M}_n U, X] \text{up}(U^T \hat{M}_n^T U) \right) \quad (104)$$

$$= - \sum_{n=1}^N \text{Tr} \left(X [U^T \hat{M}_n^T U, \text{low}(U^T \hat{M}_n U)] + X [U^T \hat{M}_n U, \text{up}(U^T \hat{M}_n^T U)] \right) \quad (105)$$

$$= \langle X, S - S^T \rangle \quad (106)$$

where we have defined $S = \sum_{n=1}^N [U^T \hat{M}_n^T U, \text{low}(U^T \hat{M}_n U)]$ and used $\sum_{n=1}^N [U^T \hat{M}_n U, \text{up}(U^T \hat{M}_n^T U)] = -S^T$. From (106) one has $\nabla \mathcal{L}(U, \mathcal{M}_\sigma) = S - S^T$ and (100) follows from the stationarity condition $\nabla \mathcal{L} = 0$. \square

Lemma 6. *Let $\mathcal{M}_\circ = \{M_n = V \text{diag}([\Lambda_{n1}, \dots, \Lambda_{nd}]) V^{-1}\}_{n=1}^N$ be a set of jointly diagonalizable matrices such that*

$$\gamma = \min_{i > i'} \sum_{n=1}^N (\Lambda_{ni} - \Lambda_{ni'})^2 > 0 \quad (107)$$

and let U_\circ be an exact triangularizer of \mathcal{M}_\circ . Then the operator

$$T = \sum_n t_n^T t_n \quad t_n = P_{\text{low}}(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n U_\circ \otimes 1) P_{\text{low}}^T \quad (108)$$

is invertible and

$$\sigma_{\min}(T) \geq \frac{\gamma}{\kappa(V)^4} \quad \|T^{-1}\| \leq \sqrt{\frac{d(d-1)}{2}} \frac{\kappa(V)^4}{\gamma} \quad (109)$$

Proof of Lemma 6 Since $U_\circ^T M_n U_\circ$ is upper-triangular for all $n = 1, \dots, N$, the matrices $(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n U_\circ \otimes 1)$ are block upper-triangular matrices and their diagonal blocks are lower triangular. For all $n = 1, \dots, N$ one has $U_\circ^T M_n U_\circ = U_\circ^T V \Lambda_n V^{-1} U_\circ$ where we have defined $\Lambda_n = \text{diag}([\Lambda_{n1}, \dots, \Lambda_{nd}])$. Then

$$t_n = P_{\text{low}}(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n U_\circ \otimes 1) P_{\text{low}} \quad (110)$$

$$= P_{\text{low}}(U_\circ^T V \otimes U_\circ^T V^{-T})(1 \otimes \Lambda_n - \Lambda_n \otimes 1)(V^{-1} U_\circ \otimes V^T U_\circ) P_{\text{low}}^T \quad (111)$$

$$= P_{\text{low}}(U_\circ^T V \otimes U_\circ^T V^{-T}) P_{\text{low}}^T P_{\text{low}}(1 \otimes \Lambda_n - \Lambda_n \otimes 1) P_{\text{low}}^T P_{\text{low}}(V^{-1} U_\circ \otimes V^T U_\circ) P_{\text{low}}^T \quad (112)$$

$$= \tilde{V} \Gamma_n \tilde{V}^{-1} \quad (113)$$

where we have defined $\Gamma_n = P_{\text{low}}(1 \otimes \Lambda_n - \Lambda_n \otimes 1) P_{\text{low}}^T$ and $\tilde{V} = P_{\text{low}}(U_\circ^T V \otimes U_\circ^T V^{-T}) P_{\text{low}}^T$, $\tilde{V}^{-1} = P_{\text{low}}(V^{-1} U_\circ \otimes V^T U_\circ) P_{\text{low}}^T$ and the last equality follows from the fact that $U_\circ^T V$ is upper triangular (see Lemma 7). The positive semi-definite matrix T can be rewritten as

$$T = \sum_{n=1}^N t_n^T t_n = W^T W \quad W = [\tilde{V}^{-T} \Gamma_1 \tilde{V}^T, \dots, \tilde{V}^{-T} \Gamma_N \tilde{V}^T]^T = (1 \otimes \tilde{V})[\Gamma_1, \dots, \Gamma_N]^T \tilde{V}^{-1} \quad (114)$$

A bound on the smallest singular value of T can be obtained as follows

$$\sigma_{\min}(T) = \sigma_{\min} \left(\tilde{V}^{-T} [\Gamma_1, \dots, \Gamma_n] (1 \otimes \tilde{V}^T) (1 \otimes \tilde{V}) [\Gamma_1, \dots, \Gamma_n]^T \tilde{V}^{-1} \right) \quad (115)$$

$$\geq \sigma_{\min}(\tilde{V}^{-1})^2 \sigma_{\min} \left([\Gamma_1, \dots, \Gamma_n] (1 \otimes \tilde{V}^T) (1 \otimes \tilde{V}) [\Gamma_1, \dots, \Gamma_n]^T \right) \quad (116)$$

$$= \sigma_{\min}(\tilde{V}^{-1})^2 \left(\min_{\|x\|=1} x^T [\Gamma_1, \dots, \Gamma_n] (1 \otimes \tilde{V}^T) (1 \otimes \tilde{V}) [\Gamma_1, \dots, \Gamma_n]^T x \right) \quad (117)$$

$$\geq \sigma_{\min}(\tilde{V}^{-1})^2 \sigma_{\min}(1 \otimes \tilde{V})^2 \left(\min_{\|x\|=1} x^T [\Gamma_1, \dots, \Gamma_n] [\Gamma_1, \dots, \Gamma_n]^T x \right) \quad (118)$$

$$\geq \sigma_{\min}(\tilde{V}^{-1})^2 \sigma_{\min}(\tilde{V})^2 \left(\min_{\|x\|=1} x^T [\Gamma_1, \dots, \Gamma_n] [\Gamma_1, \dots, \Gamma_n]^T x \right) \quad (119)$$

$$\geq \sigma_{\min}(\tilde{V}^{-1})^2 \sigma_{\min}(\tilde{V})^2 \left(\min_{\|x\|=1} x^T \text{diag} \left(\left[\sum_n [\Gamma_n]_{11}^2, \dots, \sum_n [\Gamma_n]_{\tilde{d}\tilde{d}}^2 \right] \right) x \right) \quad (120)$$

where we have defined $\tilde{d} = \frac{d(d-1)}{2}$. The minimization problem between brackets is solved by \mathbf{e}_{i_*} with $i_* = \arg \min_i \sum_{n=1}^N [\Gamma_n]_{ii}^2$ and one has

$$\gamma = \mathbf{e}_{i_*}^T \text{diag} \left(\left[\sum_n [\Gamma_n]_{11}^2, \dots, \sum_n [\Gamma_n]_{\tilde{d}\tilde{d}}^2 \right] \right) \mathbf{e}_{i_*} = \min_{j < j'} \sum_{n=1}^N (\Lambda_{nj} - \Lambda_{nj'})^2 \quad (121)$$

where i_* and $(j_*, j'_*) = \arg \min_{j < j'} \sum_{n=1}^N (\Lambda_{nj} - \Lambda_{nj'})^2$ are related by

$$i = f(j, j') \quad f(j, j') = \sum_{k=1}^{j-1} (d - k) + j' - j \quad \text{for } j < j' \quad (122)$$

This implies

$$\sigma_{\min}(T) \geq \frac{\gamma}{\kappa(V)^4} \quad \|T^{-1}\| \leq \sqrt{\frac{d(d-1)}{2}} \frac{\kappa(V)^4}{\gamma} \quad (123)$$

where we have used

$$\sigma_{\min}(\tilde{V}) = \sigma_{\min}(P_{\text{low}}(V \otimes V^{-T})P_{\text{low}}^T) \geq \sigma_{\min}(V)\sigma_{\min}(V^{-1}) = \frac{\sigma_{\min}(V)}{\sigma_{\max}(V)} = \frac{1}{\kappa(V)} \quad (124)$$

$$\sigma_{\min}(\tilde{V}^{-1}) = \sigma_{\min}(P_{\text{low}}(V^{-1} \otimes V^T)P_{\text{low}}^T) \geq \sigma_{\min}(V^{-1})\sigma_{\min}(V) = \frac{\sigma_{\min}(V)}{\sigma_{\max}(V)} = \frac{1}{\kappa(V)} \quad (125)$$

□

Lemma 7. *Let A be an upper triangular (invertible) matrix and Σ a diagonal matrix, then for any B*

$$\text{Low}(A \otimes A^{-T})(1 \otimes \Sigma - \Sigma \otimes 1)(A^{-1} \otimes A^T)\text{Low} = \text{Low}(A \otimes A^{-T})\text{Low}(1 \otimes \Sigma - \Sigma \otimes 1)\text{Low}(A^{-1} \otimes A^T)\text{Low} \quad (126)$$

Proof of Lemma 7 Let B be any matrix of the same dimension as A ,

$$M_1 = \text{mat}(\text{Low}(A \otimes A^{-T})(1 \otimes \Sigma - \Sigma \otimes 1)(A^{-1} \otimes A^T)\text{Low}) \quad (127)$$

$$= \text{mat}(\text{Low}(1 \otimes \Sigma - \Sigma \otimes 1)(A^{-1} \otimes A^T)\text{vec}(\text{low}(B))) \quad (128)$$

$$= \text{mat}(\text{Low}(A \otimes A^{-T})(1 \otimes \Sigma - \Sigma \otimes 1)\text{vec}(A^T \text{low}(B)A^{-T})) \quad (129)$$

$$= \text{mat}(\text{Low}(A \otimes A^{-T})\text{vec}([\Sigma, A^T \text{low}(B)A^{-T}])) \quad (130)$$

$$= \text{mat}(\text{Lowvec}(A^{-T}[\Sigma, A^T \text{low}(B)A^{-T}]A^T)) \quad (131)$$

$$= \text{low}(A^{-T}[\Sigma, A^T \text{low}(B)A^{-T}]A^T) \quad (132)$$

and

$$M_2 = \text{mat}(\text{Low}(A \otimes A^{-T})\text{Low}(1 \otimes \Sigma - \Sigma \otimes 1)\text{Low}(A^{-1} \otimes A^T)\text{Lowvec}(B)) \quad (133)$$

$$= \text{mat}(\text{Low}(A \otimes A^{-T})\text{Low}(1 \otimes \Sigma - \Sigma \otimes 1)\text{vec}(\text{low}(A^T \text{low}(B)A^{-T}))) \quad (134)$$

$$= \text{mat}(\text{Low}(A \otimes A^{-T})\text{Lowvec}([\Sigma, \text{low}(A^T \text{low}(B)A^{-T})])) \quad (135)$$

$$= \text{mat}(\text{Lowvec}(A^{-T}[\Sigma, \text{low}(A^T \text{low}(B)A^{-T})]A^T)) \quad (136)$$

$$= \text{low}(A^{-T} \text{low}([\Sigma, \text{low}(A^T \text{low}(B)A^{-T})]A^T)) \quad (137)$$

Then $M_1 = M_2$ can be shown by observing that $A^T \text{low}(B) A^{-T}$ is a lower-diagonal matrix if A is upper triangular. Then, for every lower-diagonal matrix C , one has

$$[\Sigma, C] = [\Sigma, \text{low}(C) + \text{diag}(C)] = [\Sigma, \text{low}(C)] = \text{low}([\Sigma, \text{low}(C)]) \quad (138)$$

because diagonal matrices always commute and the commutator of a strictly lower diagonal matrix with a diagonal matrix is strictly lower diagonal. \square

Lemma 8. *Let A be an upper triangular matrix with real nonzero eigenvalues. If A is invertible and the eigenvalues of A satisfy $\lambda_i(A) \neq \lambda_{i'}(A)$ for all $i \neq i'$ the matrix*

$$T_A = P_{\text{low}}(1 \otimes A^T - A \otimes 1)P_{\text{low}}^T \quad (139)$$

is invertible.

Proof of Lemma 8 From the spectral decomposition of the matrix A one has $A = V\Lambda V^{-1}$, with V upper triangular and Λ diagonal, and

$$T_A = P_{\text{low}}(V \otimes V^{-T})(1 \otimes \Lambda - \Lambda \otimes 1)(V^{-1} \otimes V^T)P_{\text{low}}^T \quad (140)$$

$$= P_{\text{low}}(V \otimes V^{-T})\text{Low}(1 \otimes \Lambda - \Lambda \otimes 1)\text{Low}(V^{-1} \otimes V^T)P_{\text{low}}^T \quad (141)$$

where the second equality follows from the fact that $(V^{-1} \otimes V^T)P_{\text{low}}^T \tilde{a} = \text{Low}(V^{-1} \otimes V^T)P_{\text{low}}^T \tilde{a}$ for any $\frac{d(d-1)}{2}$ -dimensional vector \tilde{a} and $(1 \otimes \Lambda - \Lambda \otimes 1)\text{Low}a = \text{Low}(1 \otimes \Lambda - \Lambda \otimes 1)\text{Low}a$ for any d -dimensional vector a . The smallest singular value of T_A obeys

$$\sigma_{\min}(T_A) \geq C_1^2 C_2 \quad (142)$$

where

$$C_1 = \sigma_{\min}(V^{-1})\sigma_{\min}(V) = \frac{\sigma_{\min}(V)}{\sigma_{\max}(V)} \quad C_2 = \min\{\|P_{\text{low}}(1 \otimes \Lambda - \Lambda \otimes 1)P_{\text{low}}^T x\|, \|x\| = 1\} = \min_{i < i'} |\lambda(A)_i - \lambda(A)_{i'}| \quad (143)$$

This implies that T_A is invertible if V is full rank and $\lambda_i(A) \neq \lambda_{i'}(A)$ for all $i \neq i'$, which are both fulfilled by assumption. \square

Lemma 9. *The Hessian of \mathcal{L} at $U = U_{\circ}e^{\alpha Y}$, where U_{\circ} is an exact triangularizer of $\mathcal{M}_{\circ} = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$ and $Y = -Y^T$, $\|Y\| = 1$, is positive definite for all Y if*

$$\alpha \leq \alpha_{\max} \quad \alpha_{\max} = \frac{2\varepsilon - \sigma A_{\sigma}}{A_{\alpha}} + O((\alpha + \sigma)^2) \quad (144)$$

$$\varepsilon = \frac{\gamma}{2\kappa(V)^4} \quad \gamma = \min_{j < j'} \sum_{n=1}^N (\Lambda_{nj} - \Lambda_{nj'})^2 \quad A_{\alpha} = 32 \sum_{n=1}^N \|M_n\|^2 \quad A_{\sigma} = 16\sqrt{N} \sqrt{\sum_{n=1}^N \|M_n\|^2} \quad (145)$$

Proof of Lemma 9 Let

$$\mathcal{L}(U, \mathcal{M}_{\sigma}) = \sum_{n=1}^N \text{Tr}(g_n^T g_n) \quad g_n = \text{low}(U^T \hat{M}_n U) \quad (146)$$

Then we have $\langle X, \nabla \mathcal{L}(U) \rangle = \frac{d}{dt} \mathcal{L}(Ue^{tX})|_{t=0} = \sum_{n=1}^N \text{Tr}(\dot{g}_n^T g_n + g_n^T \dot{g}_n)$ where $X = -X^T$ and $\dot{g}_n = \frac{d}{dt} g_n(Ue^{tX})|_{t=0} = \text{low}([U^T \hat{M}_n U, X])$. The second derivative in the direction X defines the Hessian of \mathcal{L} at U via

$$\langle X, \nabla^2 \mathcal{L} X \rangle = \frac{d^2}{dt^2} \mathcal{L}(Ue^{tX})|_{t=0} = \sum_{n=1}^N \text{Tr}(2\dot{g}_n^T \dot{g}_n + \ddot{g}_n^T g_n + g_n^T \ddot{g}_n) \quad (147)$$

where $\ddot{g}_n = \frac{d^2}{dt^2} g_n(Ue^{tX})|_{t=0} = \text{low}([U^T \hat{M}_n U, X], X)$. Let $f(U, \mathcal{M}_{\sigma})$ be a general function of $U = U_{\circ}e^{\alpha Y}$, where U_{\circ} is an exact triangularizer of $\mathcal{M}_{\circ} = \{\hat{M}_n|_{\sigma=0}\}_{n=1}^N$ and $Y = -Y^T$, $\|Y\| = 1$, and the empirical matrices \hat{M}_n . The double expansion, respect to the parameter α and σ is

$$f = f|_{(\alpha=0, \sigma=0)} + \alpha \partial_{\alpha} f|_{\sigma=0} + \sigma \partial_{\sigma} f|_{\alpha=0} + O((\alpha + \sigma)^2) \quad (148)$$

Now, consider the double expansion of the functions g_n , \dot{g}_n and \ddot{g}_n . In the first order approximation one obtains

$$\begin{aligned} \langle X, \nabla^2 \mathcal{L} X \rangle &= \sum_{n=1}^N \text{Tr} \left(2\dot{g}_n^T \dot{g}_n + \sigma(\dot{g}_n^T \partial_\sigma \dot{g}_n + \partial_\sigma \dot{g}_n^T \dot{g}_n) + \alpha(\dot{g}_n^T \partial_\alpha \dot{g}_n + \dot{g}_n^T \partial_\alpha \dot{g}_n) + \right. \\ &\quad \left. \sigma(\dot{g}_n^T \partial_\sigma g_n + \partial_\sigma g_n^T \dot{g}_n) + \alpha(\dot{g}_n^T \partial_\alpha g_n + \partial_\alpha g_n^T \dot{g}_n) \right) + O((\alpha + \sigma)^2) \end{aligned} \quad (149)$$

where the first term is always nonnegative. Now, the Hessian of \mathcal{L} at U is positive definite if $\langle X, \nabla^2 \mathcal{L} X \rangle$, for all X such that $X = -X^T$. The non negativity of (149) is guaranteed by the following condition

$$2 \sum_{n=1}^N \text{Tr}(\dot{g}_n^T \dot{g}_n) \geq \alpha \tilde{A}_\alpha + \sigma \tilde{A}_\sigma + O((\alpha + \sigma)^2) \quad (150)$$

where

$$\tilde{A}_\alpha = \left| \sum_{n=1}^N \text{Tr}(\dot{g}_n^T \partial_\alpha \dot{g}_n + \dot{g}_n^T \partial_\alpha \dot{g}_n + \ddot{g}_n^T \partial_\alpha g_n + \partial_\alpha g_n^T \ddot{g}_n) \right| \quad \tilde{A}_\sigma = \left| \sum_{n=1}^N \text{Tr}(\dot{g}_n^T \partial_\sigma \dot{g}_n + \partial_\sigma \dot{g}_n^T \dot{g}_n + \ddot{g}_n^T \partial_\sigma g_n + \partial_\sigma g_n^T \ddot{g}_n) \right| \quad (151)$$

We seek some ε , A_α and A_σ such that

$$\sum_{n=1}^N \text{Tr}(\dot{g}_n^T \dot{g}_n) \geq \varepsilon \|X\|^2 \quad A_\alpha \|X\|^2 \geq \tilde{A}_\alpha \quad A_\sigma \|X\|^2 \geq \tilde{A}_\sigma \quad (152)$$

Given ε , A_α and A_σ satisfying (152), the non negativity of the Hessian is implied by

$$2\varepsilon \geq \alpha A_\alpha + \sigma A_\sigma \quad (153)$$

from which the condition on α stated by the lemma. The explicit form of ε , A_α and A_σ are provided by Lemma 10 and Lemma 11. \square

Lemma 10. *A possible choice of $\varepsilon > 0$ satisfying (152) is given by*

$$\varepsilon = \frac{\gamma}{2\kappa(V)^4} \quad \gamma = \min_{j < j'} \sum_{n=1}^N (\Lambda_{nj} - \Lambda_{nj'})^2 \quad (154)$$

with V and Λ defined in (1).

Proof of Lemma 10 This can be seen as follows:

$$\sum_{n=1}^N \text{Tr}(\dot{g}_n^T \dot{g}_n) = \sum_{n=1}^N \text{Tr} \left(\text{low}([U_\circ^T M_n U_\circ, X])^T \text{low}([U_\circ^T M_n U_\circ, X]) \right) \quad (155)$$

$$= \sum_{n=1}^N \text{Tr} \left(\text{low}([U_\circ^T M_n U_\circ, \text{low}(X)])^T \text{low}([U_\circ^T M_n U_\circ, \text{low}(X)]) \right) \quad (156)$$

$$= \sum_{n=1}^N \text{vec}(\text{low}([U_\circ^T M_n U_\circ, \text{low}(X)]))^T \text{vec}(\text{low}([U_\circ^T M_n U_\circ, \text{low}(X)])) \quad (157)$$

$$= \sum_{n=1}^N \text{vec}(X)^T \text{Low}(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n^T U_\circ \otimes 1)^T \text{Low} \quad (158)$$

$$(1 \otimes U_\circ^T M_n^T U_\circ - U_\circ^T M_n^T U_\circ \otimes 1) \text{Lowvec}(X) \quad (159)$$

$$= \text{vec}(X)^T P_{\text{low}} \left(\sum_{n=1}^N t_n^T t_n \right) P_{\text{low}}^T \text{vec}(X) \quad (160)$$

$$= \text{vec}(X)^T P_{\text{low}} T P_{\text{low}}^T \text{vec}(X) \quad (161)$$

where we have used $\text{Low} = P_{\text{low}} P_{\text{low}}^T$ and the definition of T given in Lemma 6. For every X such that $X = -X^T$ one has $\|\text{low}(X)\| = \frac{1}{\sqrt{2}} \|X\|$. In particular

$$\text{vec}(X)^T P_{\text{low}} T P_{\text{low}} \text{vec}(X) \geq \frac{1}{2} \|X\|^2 \sigma_{\min}(T) \quad (162)$$

and using the result of Lemma 6 one obtains

$$\sum_{n=1}^N \text{Tr}(\dot{g}_n^T \dot{g}_n) \geq \frac{\gamma}{2\kappa(V)^4} \|X\|^2 \quad (163)$$

and hence $\varepsilon = \frac{\gamma}{2\kappa(V)^4}$. \square

Lemma 11. *A possible choice of A_α and A_σ satisfying (152) is given by*

$$A_\alpha = 32 \sum_{n=1}^N \|M_n\|^2 \quad A_\sigma = 16\sqrt{N} \sqrt{\sum_{n=1}^N \|M_n\|^2} \quad (164)$$

Proof of Lemma 11 Let a_α , b_α , a_σ and b_σ be defined by

$$\sum_{n=1}^N \text{Tr}(\dot{g}_n^T \partial_\alpha \dot{g}_n) = \sum_{n=1}^N \text{Tr}(\text{low}([U_\circ^T M_n U_\circ, X])^T \text{low}([U_\circ^T M_n U_\circ, Y], X)) \quad (165)$$

$$\leq \sqrt{\sum_{n=1}^N \|\text{low}([U_\circ^T M_n U_\circ, X])\|^2} \sqrt{\sum_{n=1}^N \|\text{low}([U_\circ^T M_n U_\circ, Y], X)\|^2} \quad (166)$$

$$\leq \|X\|^2 \sqrt{\sum_{n=1}^N 4\|M_n\|^2} \sqrt{\sum_{n=1}^N 16\|M_n\|^2} \quad (167)$$

$$\leq 8\|X\|^2 \sum_{n=1}^N \|M_n\|^2 \quad (168)$$

$$= a_\alpha \|X\|^2 \quad (169)$$

$$\sum_{n=1}^N \text{Tr}(\ddot{g}_n^T \partial_\alpha g_n) = \sum_{n=1}^N \text{Tr}(\text{low}([U_\circ^T M_n U_\circ, X], X)^T \text{low}([U_\circ^T M_n U_\circ, Y])) \quad (170)$$

$$\leq 8\|X\|^2 \sum_{n=1}^N \|M_n\|^2 \quad (171)$$

$$= b_\alpha \|X\|^2 \quad (172)$$

$$\sum_{n=1}^N \text{Tr}(\dot{g}_n^T \partial_\sigma \dot{g}_n) = \sum_{n=1}^N \text{Tr}(\text{low}([U_\circ^T M_n U_\circ, X])^T \text{low}([U_\circ^T W_n U_\circ, X])) \quad (173)$$

$$\leq \sqrt{\sum_{n=1}^N \|\text{low}([U_\circ^T M_n U_\circ, X])\|^2} \sqrt{\sum_{n=1}^N \|\text{low}([U_\circ^T W_n U_\circ, X])\|^2} \quad (174)$$

$$\leq \|X\|^2 \sqrt{\sum_{n=1}^N 4\|M_n\|^2} \sqrt{\sum_{n=1}^N 4\|W_n\|^2} \quad (175)$$

$$\leq 4\|X\|^2 \sqrt{N} \sqrt{\sum_{n=1}^N \|M_n\|^2} \quad (176)$$

$$= a_\sigma \|X\|^2 \quad (177)$$

$$\sum_{n=1}^N \text{Tr}(\ddot{g}_n^T \partial_\sigma g_n) = \sum_{n=1}^N \text{Tr}(\text{low}([U_\circ^T M_n U_\circ, X], X)^T \text{low}(U_\circ^T W_n U_\circ)) \quad (178)$$

$$\leq 4\|X\|^2 \sqrt{N} \sqrt{\sum_{n=1}^N \|M_n\|^2} \quad (179)$$

$$= b_\sigma \|X\|^2 \quad (180)$$

where we have defined $a_\alpha = 8 \sum_{n=1}^N \|M_n\|^2 = b_\alpha$, $a_\sigma = 4\sqrt{N} \sqrt{\sum_{n=1}^N \|M_n\|^2} = b_\sigma$, used $\|Y\| = 1$ and

$$\sum_{n=1}^N \text{Tr}(A_n B_n) = \sum_{n=1}^N \text{vec}(A_n^T)^T \text{vec}(B_n) \quad (181)$$

$$= \text{Tr}([\text{vec}(A_1^T), \dots, \text{vec}(A_N^T)]^T [\text{vec}(B_1), \dots, \text{vec}(B_N)]) \quad (182)$$

$$= \text{vec}([\text{vec}(A_1^T), \dots, \text{vec}(A_N^T)])^T \text{vec}([\text{vec}(B_1), \dots, \text{vec}(B_N)]) \quad (183)$$

$$\leq \|\text{vec}([\text{vec}(A_1^T), \dots, \text{vec}(A_N^T)])\| \|\text{vec}([\text{vec}(B_1), \dots, \text{vec}(B_N)])\| \quad (184)$$

$$= \sqrt{\sum_{n=1}^N \|\text{vec}(A_n^T)\|^2} \sqrt{\sum_{n=1}^N \|\text{vec}(B_n)\|^2} \quad (185)$$

$$= \sqrt{\sum_{n=1}^N \|A_n\|^2} \sqrt{\sum_{n=1}^N \|B_n\|^2} \quad (186)$$

Then we have

$$\tilde{A}_\alpha \leq 2\|X\|^2(a_\alpha + b_\alpha) \quad \tilde{A}_\sigma \leq 2\|X\|^2(a_\sigma + b_\sigma) \quad (187)$$

□

Lemma 12. Let U and U_\circ be respectively the approximate joint triangularizers of \mathcal{M}_σ and the exact joint triangularizer of \mathcal{M}_\circ defined in Theorem 1. For all $n = 1, \dots, N$ and all $i = 1, \dots, d$, let $\hat{\lambda}_i(\hat{M}_n) = [U^T \hat{M}_n U]_{ii}$ and $\lambda_i(M_n) = [U_\circ^T M_n U_\circ]_{ii}$. Then, for all $n = 1, \dots, N$ and all $i = 1, \dots, d$,

$$|\hat{\lambda}_i(\hat{M}_n) - \lambda_i(M_n)| \leq 2\alpha\|M_n\| + \sigma\|W_n\| + O(\alpha^2) \quad (188)$$

with α defined in Theorem 1.

Proof of Lemma 12 Let U and U_\circ be respectively the approximate joint triangularizers of \mathcal{M}_σ and the exact joint triangularizer of \mathcal{M}_\circ defined in Theorem 1. Then $U = U_\circ e^{\alpha X}$ with $X = -X^T$, $\|X\| = 1$ and $\alpha > 0$ obeying (12). Neglecting all second order terms one has

$$|\hat{\lambda}_i(\hat{M}_n) - \lambda_i(M_n)| = |[U^T \hat{M}_n U]_{ii} - [U_\circ^T M_n U_\circ]_{ii}| \quad (189)$$

$$= |[e^{-\alpha X} U_\circ^T (M_n + \sigma W_n) U_\circ e^{-\alpha X}]_{ii} - [U_\circ^T M_n U_\circ]_{ii}| \quad (190)$$

$$= |[U_\circ^T M_n U_\circ \alpha X - \alpha X U_\circ^T M_n U_\circ]_{ii} + \sigma [U_\circ^T W_n U_\circ]_{ii}| + O(\alpha^2) \quad (191)$$

$$\leq 2\alpha\|M_n\| + \sigma\|W_n\| + O(\alpha^2) \quad (192)$$

□

References

- Abed-Meraim, K. and Hua, Y. (1998). A least-squares approach to joint Schur decomposition. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, volume 4, pages 2541–2544. IEEE.
- Absil, P.-A., Mahony, R., and Sepulchre, R. (2009). *Optimization algorithms on matrix manifolds*. Princeton University Press.
- Afsari, B. (2008). Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1148–1171.
- Afsari, B. and Krishnaprasad, P. S. (2004). Some gradient based joint diagonalization methods for ICA. In *Independent Component Analysis and Blind Signal Separation*, pages 437–444. Springer.
- Anandkumar, A., Ge, R., Hsu, D., Kakade, S., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832.
- Balle, B., Quattoni, A., and Carreras, X. (2011). A spectral learning algorithm for finite state transducers. In *Machine Learning and Knowledge Discovery in Databases*, pages 156–171. Springer.

- Cardoso, J. (1994). Perturbation of joint diagonalizers. *Telecom Paris, Signal Department, Technical Report 94D023*.
- Cardoso, J.-F. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM journal on matrix analysis and applications*, 17(1):161–164.
- Colombo, N. and Vlassis, N. (2016). Tensor decomposition via joint matrix Schur decomposition. In *International Conference on Machine Learning (ICML)*.
- Corless, R. M., Gianni, P. M., and Trager, B. M. (1997). A reordered Schur factorization method for zero-dimensional polynomial systems with multiple roots. In *Proceedings of the 1997 international symposium on Symbolic and algebraic computation*, pages 133–140. ACM.
- De Lathauwer, L. (2006). A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 28(3):642–666.
- Fu, T., Jin, S., and Gao, X. (2006). Balanced simultaneous Schur decomposition for joint eigenvalue estimation. In *Communications, Circuits and Systems Proceedings, 2006 International Conference on*, volume 1, pages 356–360. IEEE.
- Haardt, M. and Nossek, J. A. (1998). Simultaneous Schur decomposition of several nonsymmetric matrices to achieve automatic pairing in multidimensional harmonic retrieval problems. *Signal Processing, IEEE Transactions on*, 46(1):161–169.
- Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press, 2nd edition.
- Konstantinov, M., Petkov, P. H., and Christov, N. (1994). Nonlocal perturbation analysis of the Schur system of a matrix. *SIAM Journal on Matrix Analysis and Applications*, 15(2):383–392.
- Kuleshov, V., Chaganty, A., and Liang, P. (2015). Tensor factorization via matrix factorization. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*.
- Pang, J.-S. (1987). A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research*, 12(3):474–484.
- Prudhomme, S., Oden, J. T., Westermann, T., Bass, J., and Botkin, M. E. (2003). Practical methods for a posteriori error estimation in engineering applications. *International Journal for Numerical Methods in Engineering*, 56(8):1193–1224.
- Sardouie, S. H., Albera, L., Shamsollahi, M. B., and Merlet, I. (2013). Canonical polyadic decomposition of complex-valued multi-way arrays based on simultaneous Schur decomposition. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 4178–4182. IEEE.